ESTUDIO DE LA POPULARIDAD DE LA UNIVERSIDAD DEL VALLE FRENTE AL PARO ESTUDIANTIL OCURRIDO ENTRE OCTUBRE Y DICIEMBRE DEL 2018 MEDIANTE EL USO DEL ANÁLISIS DE SENTIMIENTOS.

VICTOR DUVAN RUIZ OCHOA

UNIVERSIDAD DEL VALLE
INGENIERÍA DE SISTEMAS
TULUÁ
2021

ESTUDIO DE LA POPULARIDAD DE LA UNIVERSIDAD DEL VALLE FRENTE AL PARO ESTUDIANTIL OCURRIDO ENTRE OCTUBRE Y DICIEMBRE DEL 2018 MEDIANTE EL USO DEL ANÁLISIS DE SENTIMIENTOS.

Presentado por
VICTOR DUVAN RUIZ OCHOA - 201664060
victor.ochoa@correounivalle.edu.co

Supervisado por
MAURICIO LOPEZ BENITEZ, ING
mauricio.lopez@correounivalle.edu.co

PROYECTO DE FINAL DE CARRERA PARA CUMPLIR EL REQUISITO DE INGENIERÍA DE SISTEMAS

UNIVERSIDAD DEL VALLE
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
TULUÁ
2021

Trabajo de grado presentado por
Victor Duvan Ruiz Ochoa,
Como requisito parcial para la obtención del título de Ingeniero de Sistemas

Mauricio Lopez Benitez
Director

Jurado

Jurado

Resumen

El propósito de este documento es realizar una investigación sobre la popularidad, a nivel nacional, de la Universidad del Valle y si esta fue afectada por el paro estudiantil comprendido entre octubre y diciembre del 2018. Esto permitirá que la Universidad del Valle pueda conocer cuál es su imagen frente al público y de esta forma medir el impacto que un paro estudiantil tiene sobre su popularidad. En este documento se presenta el problema, los objetivos, las actividades a ser seguidas y entre otros, se plantean las conclusiones en base a los resultados obtenidos.

Palabras clave

Minería de datos, paro estudiantil, análisis del sentimiento, algoritmos de aprendizaje automático.

Abstract

The purpose of this document is to carry out an investigation on the popularity, at the national level, of the Universidad del Valle and if it was affected by the student strike between October and December 2018. This will allow the Universidad del Valle to know what it is. Their image in front of the public and thus measure the impact that a student strike has on their popularity. This document presents the problem, the objectives, the activities to be followed and among others, the conclusions are raised based on the results obtained.

Keywords

Data mining, student unemployment, sentiment analysis, machine learning algorithms.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Planteamiento del problema

1.1.1. Descripción del problema

Uno de los sucesos extraordinarios más importantes que pueda ocurrir en la universidad pública es el paro estudiantil, o cese de actividades académicas, el cual es un mecanismo que tiene la comunidad universitaria para hacer presión sobre una causa. El caso del paro ocurrido en el periodo comprendido entre octubre de 2018 y diciembre de 2018, en el cual se pretendía obtener de parte del Estado Colombiano mayores recursos para las universidades públicas, ha llevado a que muchas personas estuvieran de acuerdo con la causa de las universidades y otras tantas a disentir de esta postura.

La desinformación o el poco interés de las personas son las causas principales que hace que las personas no tengan una buena imagen de las universidades públicas pensando, tal vez, en que el paro es una pérdida de tiempo injustificada o no es más que un acto de vandalismo. Por otra parte, hay personas que conocen o se documentan acerca de las causas por la cual los estudiantes y universidades hacen ciertas demandas, mejorando su perspectiva sobre las universidades y, de cierta forma, aprobando los reclamos de la comunidad estudiantil frente a la situación actual de dichas instituciones.

La manera en que la sociedad colombiana ven a las universidades se puede ver afectada por un paro estudiantil, haciendo que las personas expresen sus emociones en la red social Twitter ya sea en beneficio o en contra de los estudiantes y universidades.

1.1.2. Formulación del problema

¿Como se ve afectada la imagen de la Universidad del Valle que perciben los usuarios de la red social Twitter frente al paro estudiantil comprendido entre octubre y diciembre 2018?.

1.2. Objetivos

1.2.1. Objetivo general

Determinar el impacto que el paro estudiantil del periodo octubre a diciembre del 2018 tuvo en la imagen de la Universidad del Valle.

1.2.2. Objetivos específicos

- Formar el conjunto de datos necesarios para la investigación, verificando la calidad de los datos, por medio de los posts hechos por las personas de la red social Twitter a nivel nacional.
- Analizar diferentes algoritmos aplicados a la minería de datos definiendo cuál es el más apropiado para la realización de la investigación.
- Adaptar una herramienta de minería de datos clasificando los datos con base a la técnica de minería de datos seleccionada determinando el sentimiento de las personas hacia la Universidad del Valle.
- Evaluar los resultados de la investigación obteniendo una conclusión acerca de la influencia que tiene un paro estudiantil frente popularidad de la Universidad del Valle.

1.2.3. Resultados esperados

Los resultados esperados asociados con cada objetivo específico se muestran a continuación en la tabla 1.1.

Tabla 1.1: Resultados esperados.

Objetivo específico	**Resultado esperado**
Formar el conjunto de datos necesarios para la investigación, verificando la calidad de los datos, por medio de los posts hechos por las personas de la red social twitter.	Data set que contenga los datos sacados de la red social Twitter, debidamente procesados y transformados velando que la calidad de los datos sea la necesaria para aplicar los métodos de minería de datos.
Analizar diferentes algoritmos aplicados a la minería de datos definiendo cuál es el más apropiado para la realización de la investigación.	Selección del algoritmo aplicado a la minería de datos más apropiado para la realización de esta investigación.
Construir una herramienta de minería de datos clasificando los datos con base a la técnica de minería de datos seleccionada determinando el sentimiento de las personas hacia la Universidad del Valle.	Implementación de una herramienta que permita clasificar y predecir los tweets de los sentimientos de las personas
Evaluar los resultados de la investigación y sacar una conclusión acerca de la popularidad de la Universidad del Valle.	Elaboración de un informe donde se especifiquen los resultados obtenidos de la investigación.

1.3. Justificación

En los últimos años las redes sociales se han vuelto muy populares entre los usuarios de internet, haciendo que cada vez más personas tengan una cuenta en la red social de preferencia. Esto ha conllevado que cada vez más haya más personas en estas redes sociales que expresan sus opiniones acerca de diferentes temas de actualidad por medio de las publicaciones que hacen a través de sus perfiles. Así mismo, muchos medios de comunicación decidieron publicar sus noticias en estas plataformas teniendo como resultado que muchos usuarios sigan sus publicaciones. Estas noticias, o incluso publicaciones de usuarios particulares, influyen en la opinión de

muchos otros usuarios haciendo que, tal vez, cambien de parecer ante diferentes temas.

Con el incremento de usuarios en las redes sociales, es importante para cualquier institución conocer la opinión que tienen estos usuarios sobre ellas, y en particular la Universidad del Valle cuando sucede un acontecimiento que puede afectar su imagen como lo es el caso de un paro estudiantil, ya que a partir de esta información se puede conocer la imagen que los usuarios de las redes sociales perciben de su institución.

Algunas noticias acerca de este paro estudiantil pudo haber echo que cierto grupo de personas cambien su visión acerca de la Universidad del Valle, ya que muchas noticias solo publican una parte del paro estudiantil y no en su totalidad, haciendo que solo se publiquen actos vandalismo y no las marchas pacíficas y las razones de las marchas, pero otros noticieros si publican la totalidad del paro estudiantil informando a las personas de las razones de porque los estudiantes y las universidades marchan. Con base en esto una persona puede cambiar su opinión acerca de una universidad ya sea por información que los medios de comunicación transmiten o por cuenta propia de la persona.

La Universidad del Valle se puede ver afectada por un paro estudiantil, ya que al afectar su imagen como institución puede influir en la decisión de las personas que están pensando hacer su formación académica en la Universidad del Valle desistir de esta decisión, o por el contrario prefiera a la Universidad del Valle en lugar de otra institución de educación superior. Por otra parte, el paro estudiantil también puede influenciar a los estudiantes de la universidad a tomar decisiones como retirarse, o abandonar sus estudios.

Así, esta investigación es un aporte a la Universidad del Valle ya que esta podrá saber cuál es su imagen ante el público y si esta imagen es afectada de alguna manera por un paro estudiantil, y con esta información la universidad podrá tomar medidas para, de alguna manera, hacer que las personas perciban una mejor imagen de la universidad.

1.4. Metodología

Para efectos del desarrollo del proyecto se planea utilizar una metodología descriptiva con la que se busca recoger, organizar, analizar, generalizar, los resultados de la investigación. Esta metodología implica la recolección de datos para dar una idea clara de los resultados de la investigación. También, esta metodología tiene

como ventaja que es fácil, a corto tiempo y económica.

Para este proyecto se cuenta con fuentes de información tanto primarias como secundarias dado que se hace uso de artículos científicos referentes al tema, libros e información de internet.

Este trabajo de grado se desarrollo con un punto de vista de investigación, por lo tanto, aunque no se implementó una metodología de desarrollo específica, el desarrollo se dividió en fases de acuerdo con los objetivos específicos, para esto se utilizó la metodología CRISP que proporciona un estándar para obtener mejores y más rápidos resultados en el proceso de la minería de datos, por esta razón, se realizó un recorrido por todos los pasos de esta metodología. Para la gestión del conocimiento de utilizó la metodología Kanban que proporciona un control claro del estado del proyecto.

1.4.1. Actividades

Tabla 1.2: Tareas del proyecto.

Objetivo específico	**Tareas**
1. Formar el conjunto de datos necesarios para la investigación, verificando la calidad de los datos, por medio de los posts hechos por las personas de la red social twitter.1.	1.1. Investigación de los métodos para acceder a los post de twitter y selección del más apropiado. 1.2. Establecimiento de cuáles serán las palabras claves y/o filtros que debe tener cada post para que sea relevante para la investigación. 1.3. Establecimiento del conjunto de datos inicial. 1.4. Realización de un debido procesado y transformación del conjunto de datos inicial.
2. Estudiar los diferentes algoritmos aplicados a la minería de datos para definir cuál es el más apropiado para la realización de la investigación.	2.1. Investigación sobre las estrategias existentes para el proceso de minería de datos. 2.2. Selección de la estrategia más conveniente para la investigación 2.3. Documentación sobre porque la estrategia seleccionada es la más apropiada.
3. Construir una herramienta de minería de datos para clasificar los datos en base a la técnica de minería de datos seleccionada para determinar el sentimiento de las personas hacia la Universidad del Valle.	3.1. Establecimiento de cuál será el subconjunto de datos para el entrenamiento del modelo. 3.2. Implementación el modelo en base a la técnica seleccionada. 3.3. Evaluación de la exactitud de las predicciones del modelo.
4. Evaluar los resultados de la investigación y sacar una conclusión acerca de la popularidad de la Universidad del Valle y si ésta se ve afectada en alguna medida por el paro estudiantil.	4.1. Elaboración de las pruebas pertinentes con los datos previamente elaborados. 4.2. Evaluación de la validez de los resultados obtenidos. 4.3. Elaboración de análisis y conclusiones con respecto a los resultados obtenidos en las pruebas.

Capítulo 2

Marco referencial

2.1. Marco teórico

2.1.1. Minería de datos

La minería de datos está conformada por una serie de herramientas y técnicas de análisis de datos que es utilizada para la extracción de información interesante por medio de la identificación de patrones, esta información puede ser utilizada como un soporte para la toma de decisiones [9]. Para el descubrimiento de conocimiento en la información se pueden utilizar varias formas de análisis, con las cuales se puede llegar a identificar patrones en los datos analizados, esta información puede ser representada a través de modelos matemáticos sobre los datos, creando un modelo de minería de datos. Luego de ser creado el modelo de minería de datos se podría analizar nueva información a través de este modelo entrenado examinando si se apega a los patrones definidos.

Origen de la minería de datos

Las diferentes técnicas de minería de datos que actualmente existen son el resultado de un largo proceso de investigaciones. Este proceso inicia cuando se empieza a almacenar la información de organizaciones de forma electrónica, después sigue con las mejoras en el acceso de los datos y por último con las nuevas tecnologías que permiten a los usuarios naveguen por la información en tiempo real. La minería de datos se basa en este proceso y va más allá del acceso a los datos, analizando la información para luego mostrar resultados. Por otra parte, las técnicas de minería de datos se empezaron a utilizar con el avance de la tecnología, ya que actualmente se cuenta con tres factores que son clave para aplicar estas técnicas: La recopilación de datos de forma masiva, computadores con cada vez más capacidad de procesamiento y el surgimiento de algoritmos más eficientes referentes a la minería de datos.

Minería de datos predictiva

La minería de datos predictiva consiste en extraer información que existe entre los datos y utilizarlos para predecir tendencias y patrones de comportamiento. La minería de datos predictiva se fundamenta en la identificación de relaciones entre variables que se presentan en eventos pasados para después aprovecharse de estas relaciones y predecir posibles resultados de futuras situaciones.

- **Clasificación:**
  Dentro de la minería de datos predictiva se utiliza la clasificación para identificar características que indican a cuál grupo, de una serie de grupos ya definidos, pertenece para cada caso, es decir, se evalúa los nuevos datos y se predice a cuál grupo tiene más probabilidad de pertenecer.

- **Regresión:**
  La regresión es utilizada para predecir o pronosticar qué valores se obtendrán en un futuro en base a los valores existentes.

2.1.2. Análisis del sentimiento

Es la técnica y el proceso de categorizar el tono emocional que existe entre una serie de palabras, oraciones o párrafos utilizado principalmente para determinar actitudes, emociones u opiniones expresadas por grupos de personas [8].

El análisis del sentimiento implica la construcción de un sistema para la recolección y el análisis de la opiniones que expresan las personas sobre algún tema en concreto, ya sea en una red social, en la publicación de blogs, reseñas o comentarios.

Actualmente existen muchas fuentes y algoritmos utilizados para analizar los sentimientos, pero este análisis tiene un cierta incertidumbre ya que el análisis del sentimiento absoluto no es posible. Según varios estudios dirigidos al análisis del sentimiento, se llegó a la conclusión de que la tasa de exactitud está entre el 70 % y el 79 %, esto de debe a diferentes factores que están implícitos en la manera como las personas expresan su opinión, entre ellas está el sarcasmo, la negación, combinaciones de palabras y la subjetividad.

2.1.3. Metodología CRISP

La metodología CRISP proporciona un ciclo de vida de un proyecto de análisis de datos, cuenta con seis fases que permiten obtener mejores resultados en un menor tiempo.

Fase 1. Comprensión del negocio: Esta es la fase inicial, se centra en la comprensión de los objetivos de proyecto y cuáles deben ser los requisitos de los resultados, para después definir un plan preliminar para alcanzar dichos objetivos.

Fase 2. Compresión de los datos: Esta fase inicia con la recopilación de los datos iniciales y prosigue con actividades para familiarizarse con los datos, identificar problemas de calidad en los datos y descubrir subconjuntos de datos interesantes y útiles para identificar información oculta entre los datos.

Fase 3. Preparación de los datos: Esta fase consta en la preparación de los datos y construcción del conjunto de datos final a partir de los datos iniciales sin procesar.

Fase 4: Modelado: En esta fase se seleccionan y aplican diversas técnicas de modelado que sean pertinentes al problema y se calibran sus parámetros a valores óptimos. Esta fase depende mucho de la fase anterior, por lo cual algunas veces se acaba volviendo a la fase de preparación de los datos para perfeccionar el modelo.

Fase 5: Evaluación: Esta fase se encarga de evaluar más a fondo el modelo construido, asegurando que se alcanza adecuadamente los objetivos planteados en la fase inicial.

Fase 6: Despliegue: En esta fase se organiza el conocimiento obtenido en la fase de modelado dependiendo de los requisitos previamente pactados, para presentarse a las personas interesadas en el proyecto, ya sea en modo de informe o en las actividades pactadas en la fase inicial.

2.1.4. Metodología Kanban

La metodología Kanban consiste en la elaboración de un cuadro o diagrama en el que se reflejan tres columnas de tareas; pendientes, en proceso o terminadas. Este cuadro debe estar al alcance de todos los miembros del equipo, evitando así la repetición de tareas o la posibilidad de que se olvide alguna de ellas. Por lo tanto, ayuda a mejorar la productividad, eficiencia del equipo de trabajo y proporciona en todo momento el estado actual del proyecto.

2.1.5. Tokenización

La tokenización es un proceso mediante el cual se convierte cada oración, párrafo, publicación, etc; en un conjunto de palabras que se conoce como tokens. En este proceso también se elimina de cada oración los tokens o palabras que no tenga

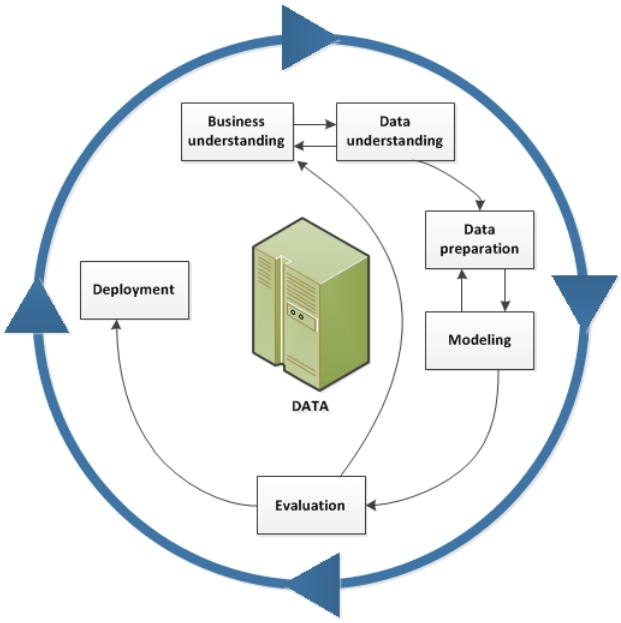


Figura 2.1: Metodología CRISP [12].

relevancia o también conocidas como las stopwords, en este sentido, las stopwords más comunes en el idioma español son los artículos, pronombres, preposiciones, conjunciones y adverbios. Otro elemento que se tiene en cuenta para la tokenización es la eliminación de los signos de puntuación, los números, los símbolos o cualquier otro elemento que no sea relevante para el análisis.

2.1.6. Steamming

Muchas veces, mas en el idioma español, diferentes palabras hacen referencia hacia un mismo concepto o conceptos muy parecidos, por ejemplo, las palabras "escribo", "escribíamos" y "escribimos" hacen referencia al mismo concepto de "escribir"; ahora, estas palabras se pueden representar por su raíz, que en este caso es "escrib". Este es el concepto de stemming, llevar cada token o palabra hacia su raíz para relacionar aquellas publicaciones que tienen tokens con similitud semántica, lo que haría mas fácil el clustering.

2.1.7. Frecuencia de términos - frecuencia inversa de documento (TF-IDF)

Esta es un técnica la cual permite representar los tokens o términos que representan una oración, publicación, párrafo, etc; de forma numérica, consiste en dos

partes, la primera es calcular la Frecuencia de terminos y la segunda es calcular frecuencia inversa de documento.

- **La frecuencia de términos (TF)**: La frecuencia de términos se obtiene a partir del número de veces que aparece cada token del vocabulario de palabras en el tweet, es decir, si la palabra se encuentra n veces en la publicación, el numero que representaría la frecuencia de esta palabra es n, luego se normalizaría entre un valor de 0 a 1. se calcula de la siguiente manera:

  $$TF = \frac{n}{N}$$

  donde n es la frecuencia calculada en la técnica anterior, y N es la cantidad de palabras totales.

  Por ejemplo, si se tiene el vocabulario:

  [“contacte”, “dario”, “henao”, “univalle”, “hicieron”, “noviembre”, “edicion”, “prueba”, “firme”, “embargo”, “posibles”, “caminos”, “podamos”, “ubicar”, “pronto”]

  la representación numérica del tweet “*Pronto contacté a Darío Henao de Univalle, lo que hicieron en noviembre fue edición de prueba, la tendremos en firme en 2019. Sin embargo, hay otros posibles caminos para que se la podamos ubicar lo más pronto.*” seria:

  [1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 1/15, 2/15].

  una desventaja de esta técnica es que no representa la importancia de cada termino dentro de todos los documentos, es decir, los términos con mas repeticiones tienden a tener una mayor importancia que aquellos que aparecen escasas veces, lo cual no representa lo que de verdad ocurre, ya que un termino que frecuenta mas en todo los documentos no posee tanta relevancia como aquellos que no son tan frecuentes.

  Para solucionar este problema, se puede combinar la frecuencia de términos (TF) con la inversa de la frecuencia con la que el término en cuestión aparece en el total de documentos (IDF).

- **Frecuencia inversa de documento (IDF)**: la frecuencia inversa de documento es un factor que ayuda a reducir la importancia de aquellos términos

que poseen una frecuencia elevada y a incrementar la importancia de aquellos términos que son pocos frecuentes. Se calcula de la siguiente manera:

$$IDF = log(\frac{D}{\{d \in D : t \in d\}})$$

donde D son el total de términos, y $\{d \in D : t \in d\}$ es el numero de documentos donde aparece el término.

Por ultimo para determinar el valor TF-IDF, puesto que es un factor, se multiplica el valor de la frecuencia de cada termino normalizado (TF), con el valor de la frecuencia inversa del documento (IDF),

$$TF - IDF = TF * IDF$$

dando como resultado una matriz donde cada columna representa las palabras o términos del vocabulario y las filas son la representación numérica TF-IDF de cada publicación.

2.1.8. Modelos de aprendizaje automático

Un modelo de aprendizaje automático es una técnica que tiene la capacidad de ser entrenado, a partir de datos que están destinados para este fin, para reconocer ciertos patrones y de esta manera predecir, clasificar y , en general, extraer conocimiento de los datos sin la necesidad de indicarle las reglas para realizar esta tarea [19].

A continuación se detallan los modelos que fueron utilizados para realizar la clasificación de las publicaciones de esta investigación.

Máquina de vectores de soporte (SVM)

Las SVM (Support Vector Machine) son un algoritmo de aprendizaje supervisado comúnmente utilizado para clasificar nuevos casos a partir de un conjunto de datos destinado a entrenar este modelo cuyos casos están clasificados previamente.

Este modelo representa los casos del conjunto de datos de entrenamiento como puntos en el espacio, y con esto los puntos que pertenecen a cualquier clase pueden ser separados por brechas entre ellos (llamados también hiperplanos) y de esta manera los nuevos puntos que este modelo predice, se le asigna la clase dependiendo del lado del hiperplano pertenece. Por otro lado, SVM permite hacer una clasificación no lineal, esto es, utilizar un diferente kernel, donde el kernel es utilizado para operar en espacios con una alta dimensionalidad y no pueden ser separados de manera lineal.

La figura 2.2 [20] representa de manera visual el espacio vectorial y los hiperplanos .

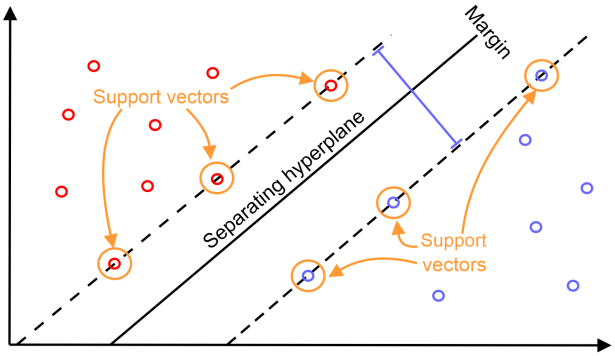


Figura 2.2: Representación gráfica de SVM.

Árboles de decisión

Este modelo de clasificación se genera de manera recursiva al dividir los datos destinados al entrenamiento del modelo, donde los nodos del árbol de decisión representan pruebas en una o más características para realizar la decisión, mientras que los nodos terminales representan la salida del modelo, es decir, la clase a la cual pertenece el caso. En la trayectoria desde la raíz hasta los nodos terminales, se identifican una combinación de características permitiendo que el modelo elija la función de cómo se divide, en cada caso, de manera que la ganancia de información al hacer la división sea lo mayor posible. En la figura 2.3 [21] se muestra visualmente en funcionamiento de este modelo con un ejemplo sobre la clasificación si se puede jugar un partido o no.

Vecinos más próximos (KNN)

KNN es uno de los modelos de clasificación más simples, a pesar de esto puede dar resultados muy competitivos en comparación con otros modelos, basa su clasificación en la búsqueda de los valores de los puntos más similares detectados en los datos de entrenamiento, para luego hacer nuevas clasificaciones basadas en estas conjeturas [22].

En comparación con otros modelos, KNN no se genera a partir de un aprendizaje con los datos de entrenamiento, sino que el propio aprendizaje ocurre en el instante en que se evalúan los datos de prueba, es decir, las nuevas clasificaciones

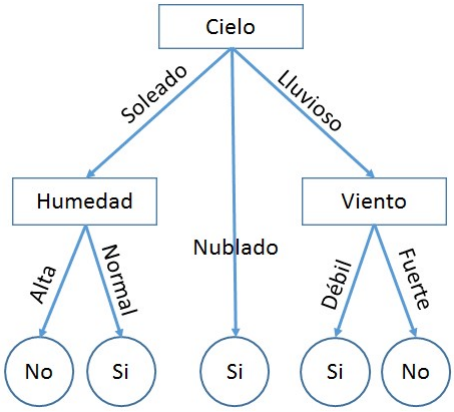


Figura 2.3: Representación gráfica de un árbol de decisión.

ocurren en el momento el que se evalúa los datos de prueba en base a los datos de entrenamiento y de los casos más cercanos en el espacio.

El funcionamiento de este modelo consiste en particionar en regiones, por localizaciones y etiquetas, los casos en los datos de entrenamiento, y de esta manera a un nuevo punto en el espacio se le asigna la clase más frecuente entre los k puntos más cercanos, la figura 2.4 [23] muestra de manera visual este proceso. Por otro lado, el parámetro k en este modelo, indica el número de casos más cercanos que se debe tener para asignarle una clase a un punto, por lo que para la implementación de este modelo es importante escoger un valor de k adecuado para tener mejores resultados en las clasificaciones.

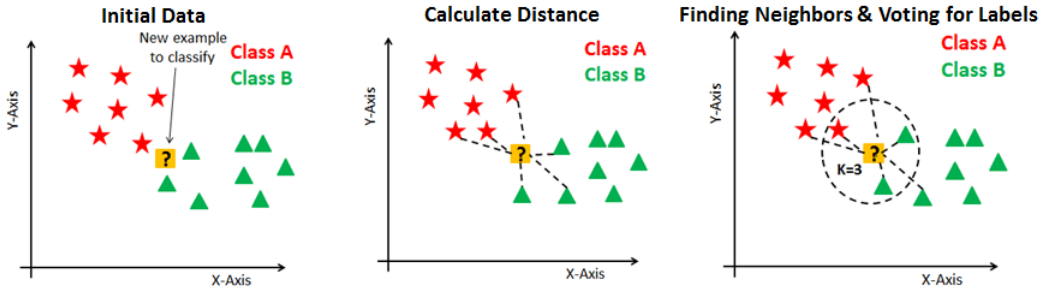


Figura 2.4: Funcionamiento del modelo KNN.

Red neuronal

Las redes neuronales son modelos simplificados que emulan el funcionamiento de las redes neuronales del cerebro humano, funciona de manera simultánea con un número de unidades de procesamiento (neuronas) que se encuentra interconectadas

unas con otras, haciendo que se parezcan a una versión abstracta de las neuronas del cerebro humano [24].

Las unidades de procesamiento, o neuronas, se organizan en normalmente en tres diferentes capas: la capa de entrada, que representan los capas de entrada, esta capa generalmente es del tamaño del total de parámetros que tengan los datos de entrenamiento; las capas ocultas (hidden layers), son un conjunto de capas de tamaños variados que unen la capa de entrada con la capa de salida; y por último la capa de salida, que representa el o los campos de salida, generalmente esta capa es de tamaño del total de clases que tenga la clasificación. Cada capa se conecta con la siguiente capa, de manera que cada neurona de la capa actual se conecta con cada una de la capa siguiente, de este modo, capa neurona de la capa de entrada se conecta con cada una de las neuronas de la primera hidden layer, si se tiene otras capas se repite esta conexión con las demás hidden layer, y por ultimo cada neurona de la última hidden layer se conecta con una de las neuronas de la capa de salida, la figura 2.5 [24] muestra una representación de un ejemplo de una red neuronal con cinco parámetros de entrada, solo una hidden layer y una salida.

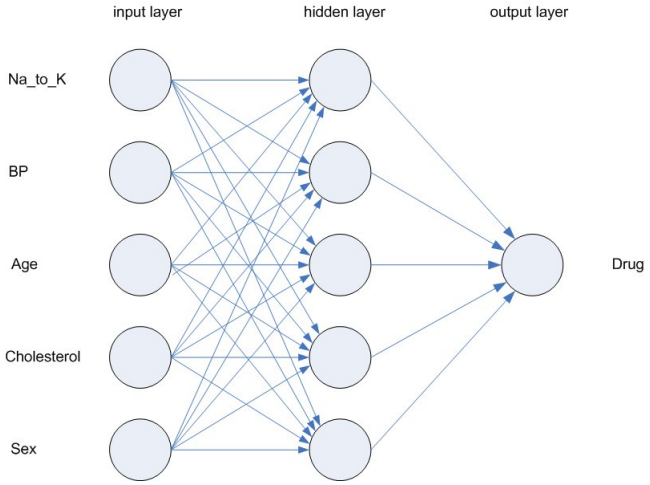


Figura 2.5: Representación gráfica de una red neuronal.

En un inicio, las estimaciones que hace la red neuronal son de manera aleatoria, pero a través de un entrenamiento con los datos destinados para ello, realiza ajustes cuando ejecuta una estimación incorrecta. Este proceso se elabora repetidas veces, haciendo que cada vez las estimaciones mejoren hasta haber alcanzado ciertos criterios de parada.

2.1.9. Técnicas de medición

Las técnicas de medición tienen como función evaluar el rendimiento de los modelos en cuanto a las clasificaciones que realicen, existen diversas métricas que permiten esta evaluación. A continuación de describirán cuatro métricas muy usadas en las evaluación de los modelos de clasificación [25].

Primero se definirán algunos parámetros que son necesarios para el entendimiento de estas métricas. En primera medida se tienen los verdaderos positivos (VP), que representan el número de muestras positivas reales que el modelo clasificó correctamente como positivos; luego están los verdaderos negativos (VN), representan el número de muestras negativas reales clasificados correctamente como negativas; después se tiene los falsos positivos (FP), que representan el número de muestras negativas reales incorrectamente clasificados como positivas; y por ultimo están los falsos negativos (FN), representan el número de muestras positivas reales incorrectamente clasificados como negativos.

Ahora se definirán las métricas que se utilizaron en esta investigación, las cuales fueron la exactitud, el recall (también conocido en español como la exhaustividad), la posición y el f1 score.

- **Exactitud:** mide el porcentaje de los casos en la cual el modelo ha acertado en la clasificación. Se calcula de la siguiente manera:

  $$Exactitud = \frac{VP+VN}{VP+VN+FP+FN}$$

  por ejemplo, si el resultado tiene una exactitud de 0.6845, quiere decir que el modelo clasificó de manera correcta el 68,45 % de los casos.

- **Recall:** esta métrica mide la proporción que el modelo identificó de manera correcta en una clase, en otras palabras responde la pregunta ¿Qué proporción de positivos reales, en una clase, se identificó correctamente?. Se calcula de la siguiente manera:

  $$Recall = \frac{VP}{VP+FN}$$

  por ejemplo, si el modelo tiene un recall en la clase "*positivo*" de 0.482, quiere decir que de los casos reales con los que se realizaron la prueba, el modelo clasifico como positivo de manera correcta el 48,2 %.

- **Precisión:** con esta métrica se mide la proporción, en una clase, que el modelo clasificó como positivo del total de elementos clasificados como positivos en dicha clase, en otras palabras responde la pregunta ¿Qué proporción de las clasificaciones positivas, en una clase, fue correcta?. Se calcula de la siguiente manera:

$$Precisin = \frac{VP}{VP+FP}$$

  por ejemplo, si el modelo evaluado tiene una precisión de 0,7623 en la clase "*negativo*", quiere decir que de los casos que se clasificaron como negativo, el 76,23 % se realizó de manera correcta.

- **F1 Score:** esta métrica es una combinación las métricas posición y recall en un solo valor, esto con el fin de hacer más fácil la comparación de los rendimientos de un mismo modelo entrenado de manera diferente, o de varios modelos. Se calcula de la siguiente manera:

$$F1Score = 2 * \frac{precision*recall}{precision+recall}$$

2.2. Marco de antecedentes

La investigación [1] propone examinar los sentimientos que tienen las personas en la red social Twitter hacia una marca, específicamente Starbucks, a través de su autenticidad, identificando las razones de los sentimientos negativos o positivos por medio de técnicas de minería de datos y análisis del sentimiento. Los resultados fueron, entre otros, que las personas no solo buscan satisfacer la necesidad del producto, sino también buscan una buena experiencia de compra, además encontraron que si una marca posee una buena autenticidad los clientes tienden a tener una mejor imagen de la marca y se inclinen más por esa marca en lugar de otras donde la autenticidad en baja o nula.

En [2] se presenta una propuesta que busca conocer los principales sentimientos de las personas de una ciudad inteligente a cerca de la gestión global del gobierno por medio de las opiniones que hacen las personas en la red social Twitter, empleando técnicas de minería de datos y análisis del sentimiento para determinar la subjetividad en términos de monitoreo positivo o negativo de dicha red social. Este trabajo se centra más en la investigación sobre actitudes, emociones y percepciones de los juicios personales. Los resultados de esta investigación varían dependiendo del tema que se trató, en la categoría de delincuencia se obtuvo una polaridad negativa,

en la categoría de trafico se obtuvo una polaridad negativa, en la categoría de finanzas públicas se obtuvo una polaridad neutral y en la categoría de town planning se obtuvo una polaridad neutral.

El estudio [3] busca, por medio del análisis del sentimiento, analizar la relación entre el nivel de aceptación en Twitter de un candidato presidencial de Ecuador y con relación a los resultados electorales. Específicamente se estudió el nivel de aceptación de las personas en la red social Twitter del expresidente de Ecuador Rafael Correa en las tres etapas de la campaña electoral, precampaña electoral, campaña electoral y postcampaña electoral. Los resultados obtenidos en cada una de las campañas electorales fueron las siguientes: Los resultados en la precampaña electoral fueron mayoritariamente una polarización negativa, con 8827 tweets negativos de 17818 tweets en total, 4264 tweets positivos y 4727 tweets neutrales. En la campaña electoral se obtuvo igualmente una polarización mayormente negativa, con 2790 tweets negativos de 5766 tweets en total, 1335 tweets positivos y 1641 tweets neutrales. En la postcampaña electoral también se obtuvo mayoritariamente una polarización negativa, con 2415 tweets negativos de 4641 tweets en total, 1096 tweets positivos y 1130 tweets neutrales. Con base en esto se concluyó que los resultados en las elecciones Ecuador 2017 no coinciden con los resultados obtenidos de este estudio, po lo cual se determinó que la relación entre las tendencias de la red social Twitter y los resultados electorales, para este caso, incluye más variables de las que actualmente fueron consideradas.

El objetivo del trabajo "Detección de tendencias en Twitter utilizando minería de datos adaptativa" [4] consiste en, mediante diferentes técnicas de minería de datos, distinguir grupos de tendencias de opiniones sobre Eurovisión, detectando posibles conexiones entre distintos países intentando predecir a través de estas conexiones un ranking de votaciones. Esta predicción se realizó con los datos obtenidos de las opiniones de las personas referentes a Eurovisión en la red social Twitter. El resultado del trabajo fue una aplicación que ayudo a la detección de tendencias en Twitter, además se concluyo que, dado que la precisión no supera el 50 % en ningún caso, al tomar solo la frecuencia de las menciones es imposible predecir resultados, ya que tuvieron en cuenta tanto las buenas como las malas opiniones de los usuarios. Comparando los resultados de las puntuaciones reales con los resultados de las puntuaciones de la predicción, se concluyó que los resultados finales fueron impredecibles aunque la aplicación predijo satisfactoriamente el candidato ganador.

El trabajo "Segmentación de Mercado Usando Técnicas de Minería de Datos en Redes Sociales" [5] tiene como objetivo en crear una aplicación de minería de datos para el descubrimiento de patrones en los datos provenientes de la red social Twit-

ter, con el fin de buscar distintas segmentaciones para ayudan a profesionales del marketing a orientar sus campañas. El resultado principal fue una aplicación capaz de segmentar diferentes publicaciones posteadas en la red social Twitter, y también el análisis de los diferentes segmentos en Twitter en el ámbito de "CapitalFederal" separando los diferentes tipos de usuarios que se pueden presentar: Usuarios tipo Bots o de creación automática, usuarios con poca actividad, usuarios que viven en un área o país a la que no apunta la campaña o usuarios con pocos seguidores.

El artículo sobre la identificacion de la opinión pública digital en la red social Twitter durante la visita del entonces candidato republicano Donald Trump a la Ciudad de México [6] busca identificar cómo está relacionada la opinión pública digital en la red social Twitter durante la visita del entonces candidato republicano Donald Trump a la Ciudad de México en el año 2016 por invitación del gobierno mexicano que fue precedida de la amenaza de construir un muro fronterizo que pagaría México. A través de métodos de minería de datos y análisis de redes sociales, el estudio busca identificar patrones de conversación y estructuras de redes que se formaron por dicho acontecimiento. Como conclusión, se pudo ratificar la influencia del contexto social y político en las conversaciones de Twitter. Se identificaron dos temas principales en dichas conversaciones, por un lado el tema de la inmigración en la contienda electoral de Estados Unidos, y por el otro usuarios indignados por los desplantes xenófobos del candidato Trump.

2.3. Marco conceptual

- **Paro estudiantil**
  Es el cese de actividades que tienen los estudiantes como mecanismo para hacer presión al estado por una causa. Consta en parar las actividades académicas de los estudiantes para realizar acciones que permitan ser escuchados por parte del estado y demandar unas necesidades.

- **Twitter**
  Es una de las redes sociales más usadas en el mundo, permite comunicarse y compartir contenidos, como fotos, vídeos u opiniones, en tiempo real desde cualquier lugar a través de internet.

- **Data set o conjunto de datos**
  Es una colección de datos almacenados, habitualmente de forma tabulada donde cada columna de la tabla corresponde a un atributo o variable en particular, y cada fila representa una instancia de ese conjunto de datos.

- **Herramienta de minería de datos**
  Son aplicaciones de software utilizadas en proceso de minería de datos que permiten clasificar cierto tipo de datos los datos, estas herramientas son entrenadas con datos particulares, por lo que son capaces de reconocer patrones en los datos y predecir nuevos datos a partir de estos patrones.

- **Web scraping**
  El web scraping es una técnica utilizada para obtener datos de una página web, consiste en obtener la página que ofrece algún sitio web y mediante software obtener los datos requeridos, como lo pueden ser un formulario, algunos párrafos o, como es en este caso, toda la información de los tweets publicados por los usuarios de Twitter.

Capítulo 3

Ingeniería del proyecto

3.1. Extracción de datos

Como primer paso para lograr los objetivos de la investigación, se inició con la recolección de datos, que para este caso fueron publicaciones hechas por los usuarios de la red social Twitter, para esto se establecieron cuales serían las características de las publicaciones, como también diversas alternativas para la extracción de los tweets pertinentes desde la propia plataforma de Twitter.

Para comenzar con la tarea de recolectar los datos que se utilizaron para el desarrollo de la investigación, primero se definió cuales serían los criterios que se deben tener en cuenta para que las publicaciones hechas por los usuarios sea consideradas como relevantes, estos fueron las características consideradas para cumplir con esto: el rango de fecha de publicación de los tweets es en las fechas en la que ocurrió el paro estudiantil, del 1 de octubre del 2018 al 31 de diciembre del 2018; las palabras clave que debe tener cada tweet son "Universidad del Valle" o "Univalle", se utilizaron estas palabras clave ya que los tweets que contengan estas palabras pueden tener alguna opinión por parte del usuario hacia la Universidad del Valle, y aunque el tweet no esté relacionado con el paro estudiantil, de alguna manera el usuario expresa su opinión hacia esta y dicho tweet se puede analizar concluyendo acerca de la imagen pública de la universidad.

También, se analizó y definió cual sería la técnica para obtener los datos. Como primera opción se intentó la obtención de los tweets mediante la API de Twitter, la cual es una herramienta brindada por la propia red social para ayudar a los desarrolladores a realizar diversas tareas que utilicen su red social, entre ellas el análisis de tweets, sin embargo esta opción fue descartada ya que la versión gratuita que ofrece esta API es muy limitada y no permite la obtención de tweets en las fechas en las que se centra esta investigación.

Una segunda opción que fue muy utilizada por trabajos de análisis del senti-

miento y en general por trabajos realizados que utilizan Twitter para obtener sus resultados, son los Firehose. Los Firehose son acuerdos que algunas compañías tienen con Twitter para obtener los tweets del usuario final, y a partir de este acuerdo los desarrolladores pueden hacer uso de estos para obtener los tweets que se necesiten. Pero, al igual que la primera opción, fue completamente descartada ya que estos acuerdos que se tenían con Twitter fueron cancelados y se dejó de brindar este servicio.

Por último, se optó por utilizar una herramienta de web scraping desarrollada en python por el autor Jefferson Henrique [17], esta herramienta utiliza el servicio de búsqueda de tweets que ofrece Twitter y mediante web scraping obtener los tweets con los parámetros de búsqueda deseados.

Como resultado se obtuvo un total de 4858 tweets con información del usuario quien publicó el tweet, la fecha exacta que fue publicado, el identificador que Twitter le asigna al tweet y el texto propio del tweet.

3.2. Preprocesamiento de datos

Con los tweets obtenidos en la tarea anterior, es necesario procesarlos permitiendo representar la información relevante como una matriz y así aplicar técnicas de minería de datos que permitan la clasificación de las publicaciones. Para esto inicialmente se filtraron algunos tweets considerados como no relevantes y se aplicaron técnicas utilizadas en el área del análisis del sentimiento, entre estas técnicas están la tekenización, el stemming y la matriz TF-IDF.

El procesamiento de las publicaciones obtenidas se realizó bajo el lenguaje de programación R, usando librerías de procesamiento de texto como los son "*stringi*", "*stringr*" o "*quanteda*".

3.2.1. Filtros

Como primera medida en el procesamiento de los datos, se filtraron los tweets de algunas cuentas de Universidad del Valle que solo proporcionaban información de carácter informativo y/o académico, por lo que no expresan ninguna opinión y no son relevantes para la investigación, estas cuentas fueron:

- XIXCAMUV, es una cuenta del congreso de actualización médica que en su mayoría de publicación solo suministraba información del tiempo del día.

- UnivalleCol, esta es la cuenta oficial de la Universidad del Valle, sus tweets son de carácter académico.
- Univalle_FSalud, esta es la cuenta oficial de la facultad de salud de la Universidad del Valle que al igual que la cuenta oficial de la Universidad solo son publicaciones de carácter académico o informativo acerca de la facultad.
- Univalle_Nic, esta es la cuenta oficial de la Universidad del Valle en Nicaragua y, aunque las publicaciones de esta cuenta podrían servir para el futuro entrenamiento del modelo, solo son publicaciones informativas acerca de la Universidad.
- univallebolivia, esta es la cuenta oficial de la Universidad del Valle en Bolivia, y ocurre exactamente el mismo caso que con la universidad en Nicaragua.
- UnivalleBun, esta es la cuenta de la sede pacifico de la Universidad del Valle, que al igual que otras cuentas oficiales de la universidad, solo son publicaciones informativas o académicas.
- univalle_pln, esta es la cuenta de un grupo de investigación y, aunque no tiene muchas publicaciones, se descarto por que las publicaciones no tienen relevancia para esta investigación.

También se eliminaron los tweets que por alguna razón se hubieran publicado varias veces ya sea por la misma cuenta en diferentes fechas o por cuentas diferente, esto con el fin de tener variedad en los datos y no tener las mismas instancias en los datos que de alguna manera pudiera afectar la muestra para el entrenamiento del futuro modelo de clasificación.

3.2.2. Muestra y clasificación de las publicaciones

Con el fin de crear el conjunto de datos de entrenamiento para los modelos de clasificación, se escogió un conjunto publicaciones como muestra, los cuales fueron clasificados manualmente a través de una aplicación diseñada para ese fin.

El proceso de clasificación consistió en leer cada una de las publicaciones de la muestra, determinando si lo que quiere decir su autor a través de dicha publicación es de carácter negativo, positivo o neutro. Este proceso se realizó un total de tres veces corrigiendo la posibilidad de que la publicación no haya sido clasificada correctamente. En el caso de que no se llegara a un acuerdo de la correcta clasificación de una publicación en particular, se pidió la opinión de dos o tres personas "naturales" para así llegar un acuerdo y lograr que la clasificación manual de las publicaciones

sea lo más correcta posible.

La aplicación que se creó para este fin consiste en una pequeña interfaz gráfica donde se podía leer la publicación en cuestión y el autor que lo público, al igual de la opción de clasificar el sentimiento como positivo, negativo o neutro.

Con esto, se obtuvo una muestra de 900 publicaciones clasificadas listas para entrenar los modelos que se implementaran, después de un previo procesamiento que se describirá en este capítulo.

3.2.3. Tokenización

Para esta investigación se realizó el proceso de tokenización en el que cada tweet se convirtió en un conjunto de palabras, removiendo las stopwords en español, eliminando la tildes y eliminando los signos de puntuación y símbolos, eliminando los números y todo carácter que no fuera alfanumérico; también se convirtió cada palabra en minúscula para evitar redundancia con la misma palabra que tuviera alguna letra en mayúscula, y se ignoraron algunas palabras que se encontraron que no tiene mayor relevancia para el análisis, la cuales fueron: “www”, “http”, “https”, “com”, “html”; y descartando también los token con una longitud menor que 3, ya que una palabra de dos letras o menos no puede ser relevante y puede aportar ruido a la investigación, y los tokens con una longitud mayor que 15, ya que son considerados como agrupaciones de letras sin sentido.

Esto deja como resultado que cada tweet que se tiene en el conjunto de datos se convierta en un conjunto solo con las palabras que se consideran relevantes y con las cuales se continuará realizando el procesamiento de los datos.

3.2.4. Stemming

El proceso de steamming consiste en convertir cada palabra que se tiene en el diccionario, en su raíz la cual representa esta palabra. En un inicio se consideró implementar este proceso con los datos de esta investigación, pero a pesar de que esta herramienta pueda ser útil, pueden surgir algunos errores en el proceso que pueden perjudicar el resultado; uno de estos posibles errores es que dos o más palabras con significados diferentes se reduzcan a la misma raíz, lo cual no debería ser así, por ejemplo, las palabras “casa” y “casamiento” se predirán reducir a una misma raíz “cas”. Otro posible error que pueda surgir es cuando dos o más palabras se deberían reducir a una misma raíz, pero se hacen a raíces diferentes y se consideran que provienen de diferentes palabras.

Por esto y por los resultados obtenidos en los experimentos de [18] sobre como el stemming afecta en los resultados, se optó por no implementar la técnica de stemming en el procesado de datos, y dejar los tokens en su estado natural.

3.2.5. Frecuencia de términos - frecuencia inversa de documento (TF-IDF)

Una vez las palabras fueron procesadas y tokenizadas, se generó un vocabulario que servirá como las características de la matriz que se ira generando más adelante. Con esto, se puede considerar cada palabra del vocabulario seleccionado como una columna de la matriz y las filas serán la representación numérica de cada publicación.

Con el fin de representar el conjunto de datos resultante hasta ahora de forma numérica, se implementó la técnica TF-IDF que hace una representación numérica de los términos de forma detallada del cual se puede obtener mejores resultados.

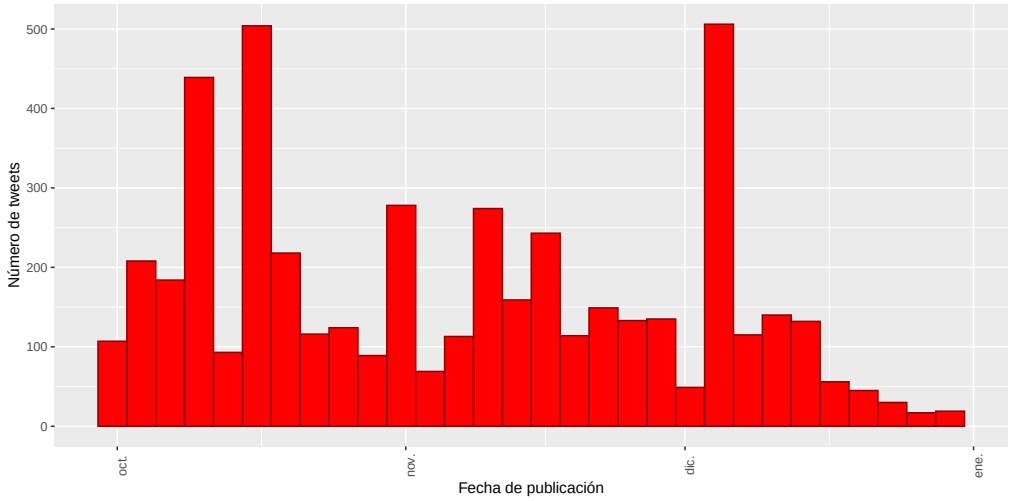
Por último, se realizó un ajuste de los términos dispersos de todo el vocabulario, esto es, ignorar los términos que tienen una frecuencia de documento inferior a un umbral determinado, esto ayuda a la generalización, es decir, que haya más términos generales en todos los documentos el lugar de términos que solo aparezcan pocas veces en pocos documentos, y de cierta manera eliminar aún más el ruido que se pueda producir a la hora de implementar el modelo. En este caso el mínimo de documentos en el cual el termino tiene que aparecer es 9, dejando un total de un poco más del 97 % de dispersión. Esto deja como resultado una matriz con 200 columnas, que representan los términos después de todo el procesamiento realizado, y 900 filas que representan las publicaciones de muestra seleccionadas.

3.3. Análisis exploratorio de los datos

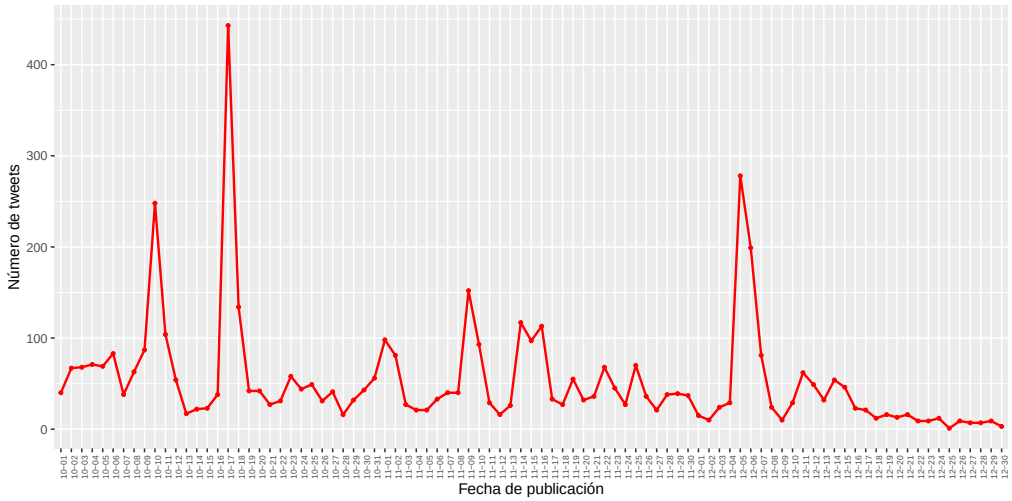
El análisis exploratorio de los datos permite examinar un poco mejor como están estructurados los datos, observando características o patrones que se puedan estar presentes en ellos. A continuación se presentara una serie de gráficos y análisis de ciertas características que se identificaron en las publicaciones obtenidas.

La primera característica que se identificó es la distribución de las publicaciones a través del tiempo, para esto en la figura 3.1a se muestra la cantidad de publicaciones que se han hecho a lo largo del tiempo en la que se centra la investigación. Para ver esto más a detalle, la figura 3.1b muestra la cantidad de publicaciones

hechas cada día desde el primero de octubre hasta el 30 de diciembre, se observa que hay varios picos donde la cantidad de publicaciones hechas superen las 100 pero en general la media de publicaciones por día fueron 53, el día con mayor número de publicaciones fue el 16 de octubre con un total de 443 y el menor número de publicaciones ocurrió el día 25 de diciembre donde solo se realizó 1 publicación, es decir, no hubieron publicaciones.



(a) Histograma: Número de tweets publicados en el tiempo.



(b) Número de tweets publicados en el tiempo.

Figura 3.1: Número de tweets publicados en el tiempo.

Otro parámetro interesante que se obtuvo fue las palabras, dentro del vocabulario obtenido, que fueron más utilizadas en las publicaciones por los usuarios, en la figura 3.2 se muestran las cinco palabras más frecuentes, sin contar las palabras que fueron criterio de búsqueda y por ende se encuentran en la gran mayoría de las publicaciones ("*univalle*", "*universidad*", "*valle*"); en primer lugar esta la palabra "*estudiantes*" con un total de 948 repeticiones, en segundo lugar la palabra "cali" con 743, en tercer lugar la palabra "educacion" con 438 repeticiones, luego le sigue la palabra "facebook" con 430 y por último la palabra "publica" con 416 repeticiones.

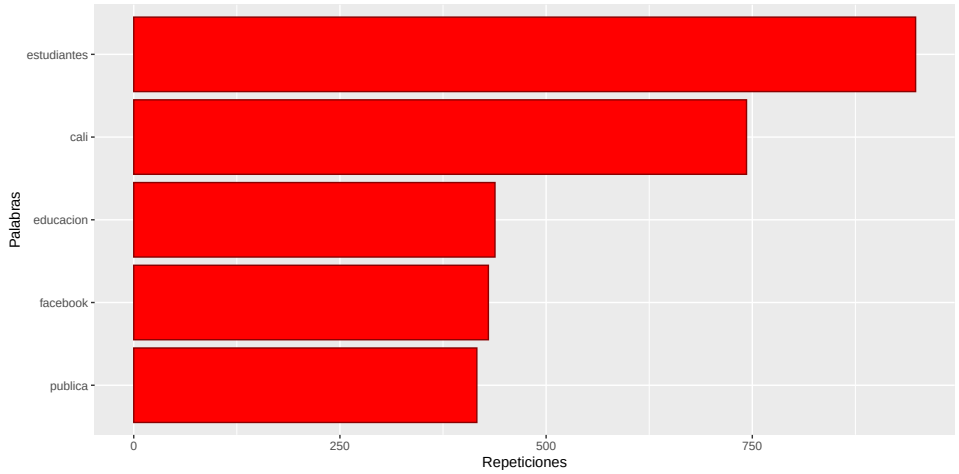


Figura 3.2: Palabras más usadas.

Como tercera característica se identificaron los usuarios con mayor número de publicaciones realizadas, la figura 3.3 muestran los cinco primeros usuarios; el usuario con más publicaciones fue "*UVsocioeconomia*" con un total de 77, en segundo lugar es el usuario "*julietacomunica*" con 48 publicaciones, el tercer usuario con más publicaciones es "*JorgeOvalleB*" con 45, seguido por "*kuijii*" con 43 publicaciones y por último el usuario "*EnterateCali*" con 35 publicaciones.

3.4. Implementación del modelo

Una vez que se obtiene la matriz con la representación numérica de las publicaciones de muestra, con su respectiva clasificación del sentimiento, se continuo con la tarea de entrenar algunos modelos de clasificación, con el fin observar el comportamiento de cada uno y obtener cuál de ellos tiene mejores resultados, para posteriormente clasificar todas las publicaciones recolectadas y poder analizar resultados que se obtengan. Para esto se implementaron cuatro modelos de clasificación,

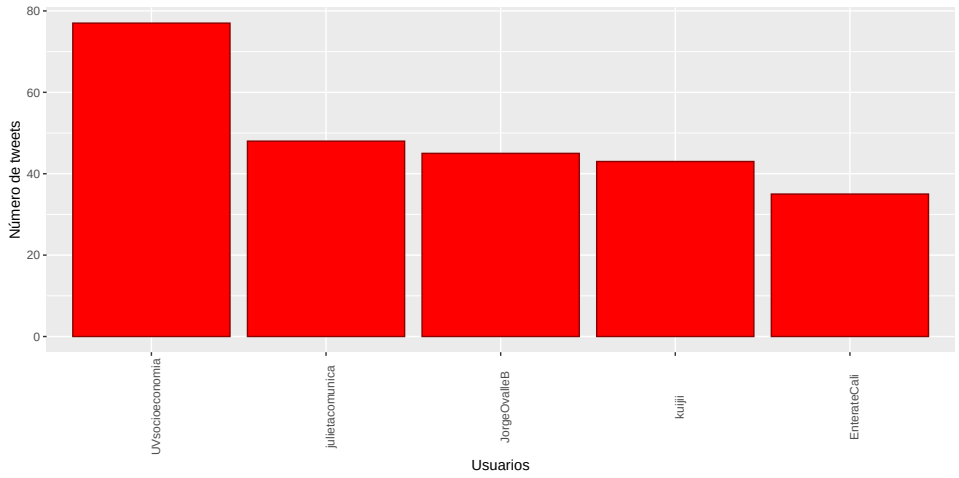


Figura 3.3: Usuarios con más publicaciones.

los cuales se variaron los parámetros correspondientes a cada modelo permitiendo que el modelo se comporte de diferentes maneras y observar cuál de estos comportamientos es más conveniente.

Al igual que el procesamiento de datos, los modelos de clasificación se implementaron en el lenguaje R usando librerías como "*e1071*", "*naivebayes*" o "*rpart*" que permiten crear estos modelos.

Se opto por separar los datos de muestra, que están destinados para el entrenamiento de los modelos, de la siguiente manera: un 70 % de los datos para entrenamiento y un 30 % para prueba.

3.4.1. SVM

El primer modelo de clasificación que se implementó con el fin de ver el comportamiento con los datos de la investigación fue la máquina de vectores de soporte. La implementación se realizó mediante la librería `e1071` la cual permite la creación de modelos SVM con diferentes kernels, costos y distintas variaciones de parámetros.

Para determinar cuál es la configuración de los parámetros con la que se tiene mejores resultados, se realizaron diferentes pruebas con los distintos parámetros del modelo, la figura 3.4 muestra como fue el proceso para determinar estos parámetros.

Figura 3.4: Proceso para determinar la configuración ideal del modelo SVM.

Las pruebas consistieron entrenar el modelo con los diferentes kernels: `linear`, `polynomial`, `radial` y `sigmoid`; observando con cual se obtiene mejores resultados. Luego se modificó el costo en un rango de 1 a 5 comparando cada combinación, para de esta manera determinar cuál es la mejor configuración de parámetros para el conjunto de datos de entrenamiento de la investigación. Se observo que la combinación de parámetros con la que se obtuvo mejores resultados fue con el kernel `sigmoid` y con un costo de 1.

3.4.2. Arboles de decisión

En segundo lugar, el modelo de clasificación escogido fue arboles de decisión, se implementó con ayuda de la librería `rpart`. Se opto por dejar el método (el parámetro `method`) por defecto con el fin de que la misma rutina haga la suposición de manera inteligente.

3.4.3. KNN

Otro modelo implementado es el k vecinos más próximos, se realizó con la ayuda de la librería `class` que permite crear modelo con diferentes valores de k. Para determinar cuál es el parámetro k con el cual se obtiene mejores resultados, se realizaron varias pruebas modificando el parámetro k, la figura 3.5 muestra el procedimiento para determinar el mejor valor de este parámetro.

Figura 3.5: Proceso para determinar la configuración ideal del modelo KNN.

Las pruebas consistieron, en primer lugar, en determinar un valor inicial de k igual a 1 y evaluar el modelo con este valor, luego se varió el parámetro en un rango de 1 a 5 evaluando y comparando los resultados del modelo en cada caso.

Como resultado se determinó que el parámetro más adecuado o con el que mejores resultados se obtiene es con $k = 3$.

3.4.4. Red neuronal

Por último, el modelo implementado para la clasificación de las publicaciones fue una red neuronal, esto realizó mediante la librería `neuralnet` que permite crear redes neuronales con diferentes configuraciones. Para determinar cuál es la configuración, o el número y el tamaño de cada hidden layer de la red neuronal con la que mejor resultados se obtienen, se evaluaron diversas combinaciones, la figura 3.6 muestra el proceso para determinar esta configuración ideal.

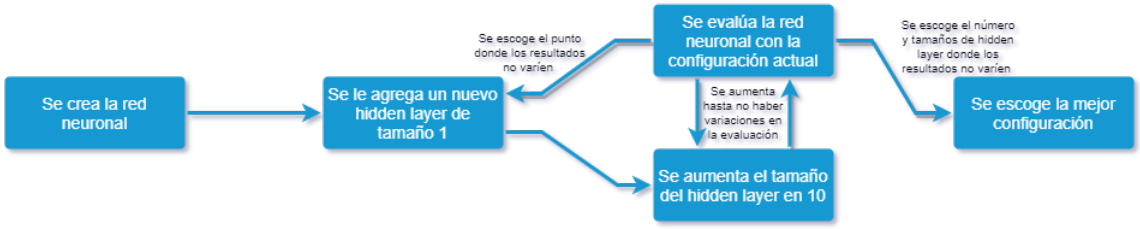


Figura 3.6: Proceso para determinar la configuración ideal de la red neuronal.

El procedimiento consistió, en primer lugar, en determinar cuál es el número de hidden layer con el que la red neuronal clasifica de mejor manera, para esto se implementó un numero de hidden layer entre 1 y 5, en el caso de cuando se implementa un solo hidden layer, el tamaño de este se fue incrementando hasta que las variaciones en los resultados no fueran relevantes, luego cuando se agrega un nuevo hidden layer, también fue incrementando como en el primer caso; y de esta manera hasta implementar los 5 hidden layer. Se observo que a partir de cierto número de hidden layer los resultados que se obtenían no variaban considerablemente, o en algunos casos empeoraban, por lo que se determinó que la configuración de la red neuronal con la que mejor resultados se obtenían era con 3 hidden layer cada uno de tamaño 150, 40, 20 correspondientemente. Por otra parte, se comprobó si al modificar el umbral (threshold) hubiera una mejora en los resultados obtenidos por la red neuronal, y se observó que al modificar el valor por defecto de 0,01 no hubo una mejoría en los resultados y en algunos casos los empeora, por lo que opto por dejar este valor por defecto.

Capítulo 4

Evaluación y Análisis de resultados

4.1. Evaluación

Una vez configurados los modelos con los respectivos parámetros establecidos, se evaluaron sus resultados después del entrenamiento; para esto se utilizaron cuatro métricas las cuales, por medio de las clasificaciones de los datos de prueba, permiten medir el desempeño de los modelos, las métricas utilizadas fueron: la exactitud, que mide la proporción de aciertos que tuvo el modelo; el recall, el cual mide la proporción de aciertos reales en cada clase que tuvo el modelo; la precisión, la cual mide que proporción de los aciertos que tuvo el modelo en cada clase se hizo correctamente; y por último el f1 score, que es una combinación de las métricas recall y precisión.

4.1.1. Resultados sin técnicas de balanceo de datos

La etapa de entrenamiento se lleva a cabo con una porción del dataset equivalente al 70 % del total de datos; dicho entrenamiento se aplicó en cada modelo con los parámetros que se establecieron anteriormente, luego, para lograr medir el desempeño del modelo, se utilizaron los datos destinados para pruebas, el 30 % de los datos restantes, con el fin de clasificarlos, y de esta manera comparar, mediante las métricas de exactitud y f1 score (ya que f1 score es una combinación de las métricas recall y posición), las clasificaciones que se obtuvieron a través de los modelos y las clasificaciones reales. En la tabla 4.1 se observan los resultados que se obtuvieron.

Haciendo un análisis a los resultados obtenidos, se observa que el modelo SVM tiene la mayor exactitud en las clasificaciones en comparación con los demás modelos, por otro lado si se observa la métrica f1 score, la clase *neutro* cuenta con un buen porcentaje de aciertos (un 85 %), pero en la clase *positiva* este porcentaje es muy bajo (un 12 %), y también la clase *negativa* cuenta con bajo porcentaje (un 36 %); esto sucede de manera similar con los resultados de los demás modelos de

Tabla 4.1: Porcentaje de la medidas de los modelos implementados con datos sin técnicas de balanceo.

	Exact	**F1 Score**			**Recall**			**Precisión**		
		Neg	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**
SVM	74,3	32,78	85,01	13,33	25	94,53	7,69	47,61	77,23	50
Árbol	69,88	36,66	81,42	13,33	27,5	87,43	11,53	55	76,19	15,78
KNN	71,08	15,68	82,52	17,14	10	92,89	11,53	36,36	74,23	33,33
NN	62,65	23,37	76,16	28,57	22,5	75,95	30,76	24,32	76,37	26,66

clasificación.

Al examinar los datos de entrenamiento y pruebas, se determinó que se obtuvo puntuaciones tan altas en una clase y baja en otras en la puntuación de f1 score, debido al desbalanceo de casos en los datos de entrenamiento, ya se contaba con 145 casos en la clase negativa, 637 casos en la clase neutro y 118 casos en la clase positiva. Para intentar corregir este problema, e intentar nivelar los casos en cada clase, se implementó tres técnicas de balanceo de datos: dawnsampling, upsampling y smote.

4.1.2. Dawnsampling

El downsampling es una técnica de balanceo de datos que consiste en descartar de manera aleatoria datos de la clases las cuales tienen la mayoría de los casos, balanceando el conjunto de datos de manera descendente, es decir, llevar las demás clases al tamaño de la clase con menor numero de casos en el conjunto de datos.

Una vez aplicada esta técnica, el conjunto de datos destinado al entrenamiento y a las pruebas del modelo quedó con un tamaño de 118 casos en cada clase. En la tabla 4.2 se muestran los resultados que se obtuvieron después de entrenar y realizar las pruebas a cada modelo con esta técnica.

Tabla 4.2: Porcentaje de la medidas de los modelos implementados con downsampling.

	Exact	**F1 Score**			**Recall**			**Precisión**		
		Neg	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**
SVM	50	52,38	56,41	33,33	42,30	84,61	23,07	68,75	42,30	60
Árbol	47,44	54,54	51,35	31,57	46,15	73,07	23,07	66,66	39,58	50
KNN	44,87	19,35	43,63	57,14	11,53	46,15	76,92	60	41,37	45,45
NN	50	53,57	50,90	44,44	57,69	53,84	38,46	50	48,27	52,63

Al observar los resultados, se hace notable la disminución de la exactitud de los modelos de clasificación, ya que, por ejemplo, el modelo SVM cuando no se aplicó ninguna técnica de balanceo de datos contaba con una exactitud del 74,3 % y en este caso cuenta con 50 %, por otro lado, al observar el f1 score se nota que las predicciones un poco más balanceadas. En este caso, los modelos SVM y red neuronal parecieran obtener los mejores resultados al aplicar esta técnica a los datos de entrenamiento.

A pesar de haber una posible disminución en los resultados de los modelos al aplicar esta técnica de balanceo de datos, estos resultados se tuvieron en cuenta, al igual que los demás resultados, para obtener el modelo definitivo el cual se utilizó para obtener las conclusiones de la investigación.

4.1.3. Upsampling

El upsampling es una técnica de balanceo de datos parecida a la técnica anterior, con la diferencia que esta técnica duplica de manera aleatoria los casos de las clases minoristas, de manera que las demás clases quedan del tamaño de la clase mayoritaria. Después de aplicar esta técnica, se obtuvieron en el conjunto de datos un total de 637 casos en cada clase.

En la tabla 4.3 se observan los resultados que se obtuvieron al aplicar esta técnica a los datos, entrenar los modelos de clasificación y evaluarlos.

Tabla 4.3: Porcentaje de la medidas de los modelos implementados con upsampling.

	Exact	**F1 Score**			**Recall**			**Precisión**		
		Neg	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**
SVM	59,2	59,13	62,35	56,21	60,10	60,65	56,83	58,20	64,16	55,61
Árbol	44,44	41,83	52,41	30,70	34,97	77,04	21,31	52,03	39,71	54,92
KNN	42,44	43,36	49,21	28,88	36,61	68,85	21,85	53,17	38,29	42,55
NN	48,09	38,35	53,59	48,10	30,60	75,40	38,25	51,37	41,56	64,81

Se observan en los resultados una ligera mejora en el balanceo en las clases clasificadas en comparación con el caso de los datos sin técnicas de balanceo de datos, también, en comparación con la técnica anterior, se ve que obtuvieron resultados similares, en el caso del modelo SVM se notó una mejoría en tanto la exactitud como en el f1 score.

4.1.4. Smote

El smote (Técnica sintética de sobremuestreo de minorías) es una técnica de balanceo de datos basada en KNN para crear de manera sintética nuevos casos con el fin de balancear las clases minoristas del conjunto de datos, en otras palabras, se crea sintéticamente casos, a partir de los ya existentes, para aumentar el número de casos de las clases minoritarias con el objetivo balancear las clases. Se aplicó esta técnica con el conjunto de datos de entrenamiento y se obtuvo en la clase negativo 75 casos, en la clase neutro 397 casos y en la clase positivo 354 casos. En la figura 4.4 se muestran los resultados que se obtuvieron luego de entrenar y evaluar los modelos después de aplicar esta técnica de balanceo de datos.

Tabla 4.4: Porcentaje de la medidas de los modelos implementados con smote.

	Exact	**F1 Score**			**Recall**			**Precisión**		
		Neg	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**
SVM	47,8	0	64,68	0	0	100	0	0	47,80	0
Árbol	54,95	13,79	67,60	42,62	11,76	82,75	33,33	16,66	57,14	59,09
KNN	60,44	19,04	58,50	66,32	11,76	49,42	83,33	50	71,66	55,08
NN	62,09	30,76	66,29	62,5	23,52	67,81	64,10	44,44	64,83	60,97

En los resultados se observa, en el caso de la exactitud y excluyendo el modelo SVM, una mejora en el porcentaje ya que en el caso de los modelos KNN Y NN se superó el 60 %, por otra parte, observando f1 score se observa que en la clase negativa el porcentaje es bajo como por ejemplo los modelos de árbol de decisión y KNN no llegan al 20 %. En el caso del modelo SVM, no fue capaz de clasificar casos en las clases negativo y positivo por lo que hace que tenga una exactitud tan baja del 47 %.

Al analizar los resultados que se obtuvieron de cada modelo con los diferentes métodos de balanceo de datos, se hace claro la disminución en la exactitud de cada uno en comparación cuando se entrenaron estos modelos sin ninguna técnica, por otro lado, hubo una mejora en cuanto a el balanceo de las clases clasificadas por los modelos gracias a este balanceo de datos. Al analizar cada uno de los resultados se tomó como posibles candidatos, para escocer el modelo definitivo que se utilizará para la clasificación de las publicaciones, los modelos SVM aplicando la técnica de upsampling y los modelos KNN y NN con la técnica smote. Se optó por el modelo SVM con la técnica de upsampling ya que, aunque la exactitud no es tan alta como en los otros candidatos, es muy cercana a estos, y adicionalmente, la clasificación en cada clase es notablemente mejor, superando más del 55 % en cada caso.

4.2. Análisis de resultados

Una vez definido y entrenado el modelo con el cual se clasificaron los datos, como siguiente paso, se clasificaron el total de publicaciones que se obtuvieron. Para esto se aplicaron las mismas técnicas para la representación numérica de las publicaciones, es decir, se aplicó el proceso de tokenización, la frecuencia de términos y la frecuencia inversa de documento; luego, haciendo uso del modelo seleccionado, se clasificaron las 4858 publicaciones obtenidas de Twitter, obteniendo los siguientes resultados.

La figura 4.1 se muestra la distribución de clasificaciones obtenidas, en las que se clasificó un total de 1189 publicaciones como negativas, 2457 como neutras y 1212 como positivas.

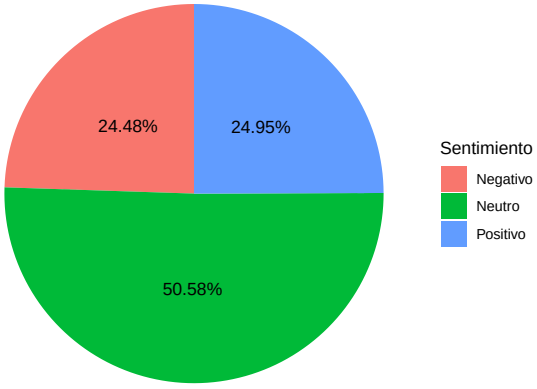


Figura 4.1: Porcentaje de las clasificaciones de las publicaciones.

El análisis realizado a partir de las clasificaciones que se obtuvieron, permite evidenciar ciertas características en los datos, una de ellas se observa en la figura 4.2, donde se puede ver la cantidad de publicaciones hechas a través del intervalo de tiempo donde se centró la investigación. En el gráfico se nota tres picos donde fueron los días donde las usuarios realizaron mayor número de publicaciones, el primer pico es el día 10 de octubre con 36 publicaciones negativas, 119 publicaciones neutras y 93 publicaciones positivas; el segundo pico ocurrió el día 17 de octubre con 204 publicaciones negativos, 157 publicaciones neutras y 82 publicaciones positivas; y el ultimo pico ocurrió el día 5 de diciembre con 160 publicaciones negativas, 79 publicaciones neutras y 36 publicaciones positivas.

También, se identificaron las palabras más frecuentes empleadas por las personas que realizaron las publicaciones en cada sentimiento, en la figura 4.3 se puede

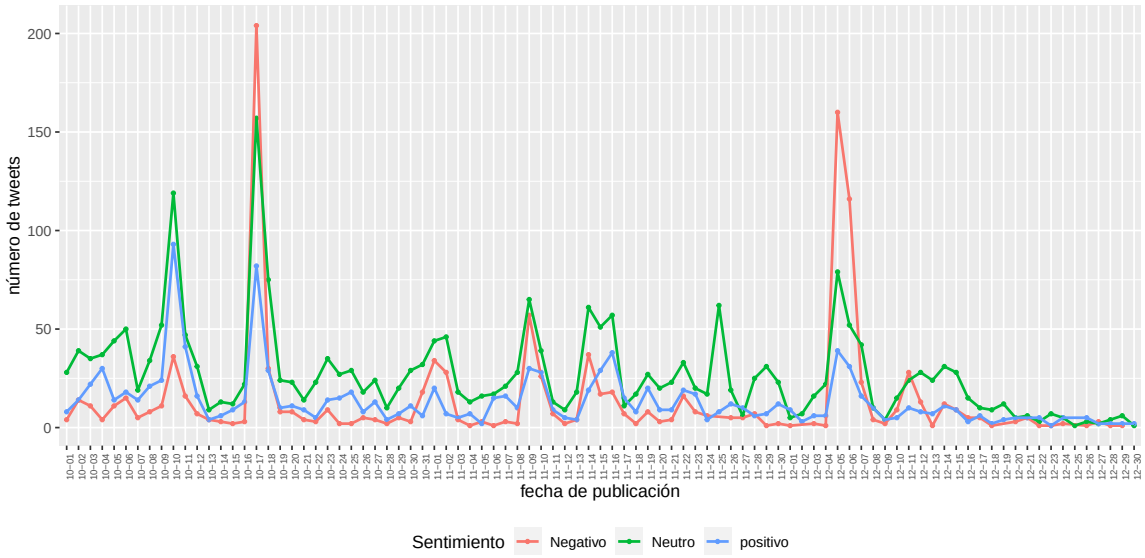


Figura 4.2: Numero de tweets publicados por sentimiento.

evidenciar estas palabras. Las palabras “estudiantes” y ”cali” obtuvieron la primera posición en los sentimientos negativo y neutro con 392 y 346 correspondientemente en el sentimiento negativo, y 346 y 413 correspondientemente en el sentimiento neutro; en el sentimiento positivo las palabras más usadas fueron “educacion” con 254 repeticiones y “publica” con 252.

La última característica que se identificó corresponde a los usuarios que más publicaciones realizaron por cada sentimiento, la figura **??** muestra estos usuarios. En el sentimiento negativo, el usuario que más realizó publicaciones fue “EnterateCali” con 20 publicaciones, seguido del usuario “elpaiscali” con 18; en el sentimiento neutro el usuario con más publicaciones fue “UVsocioeconomia” con 60 publicaciones y en segundo lugar el usuario “julietacomunica” con 48; y en el sentimiento positivo también está en primer lugar el usuario “UVsocioeconomia” con 15 publicaciones, seguido por el usuario “LFCaicedoMANE” con 8 publicaciones.

4.2.1. Clasificación de nuevos casos

Haciendo uso del modelo clasificador escogido, se implementó un prototipo web el cual permite clasificar nuevas opiniones que se puedan presentar. Para esto se hizo uso de la herramienta de R, `Shiny`, destinada al desarrollo de aplicaciones web iterativas.

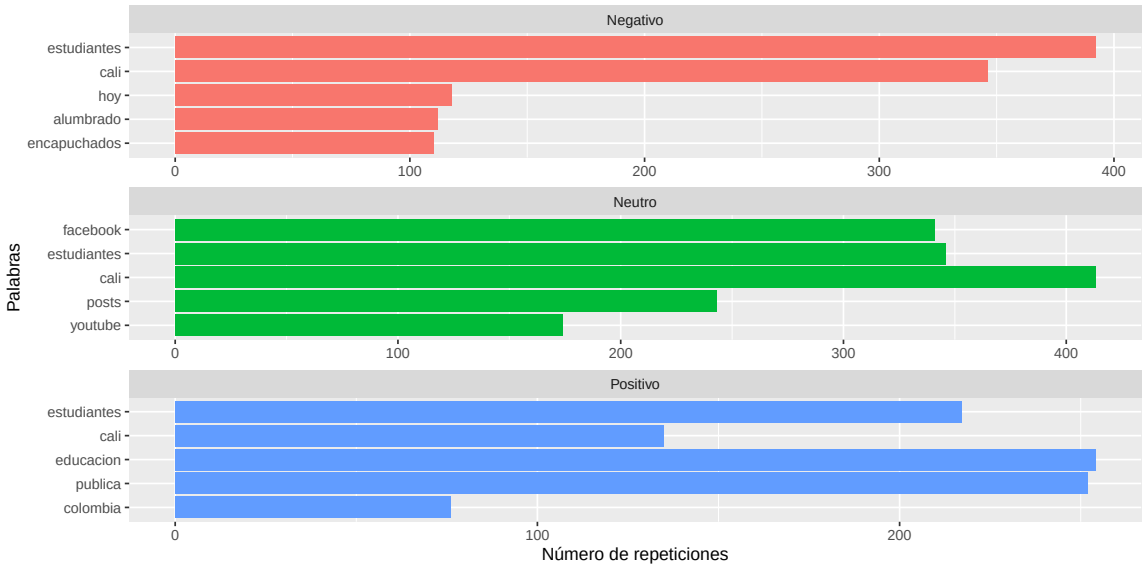


Figura 4.3: Palabras mas usadas por sentimiento.

Este prototipo se realizó de manera que al ingresar la opinión u oración, a esta se le realiza un preprocesamiento para tokenizar las palabras y representarlas numéricamente mediante TF-IDF, luego, haciendo uso del modelo seleccionado (SVM entrenado con la técnica de balanceo de datos upsampling) se clasifica y se visualiza en pantalla esta clasificación.

Se uso de la plataforma `shinyapps.io`, la cual permite alojar aplicaciones web implementadas en shiny. Se puede encontrar el prototipo web en [26].

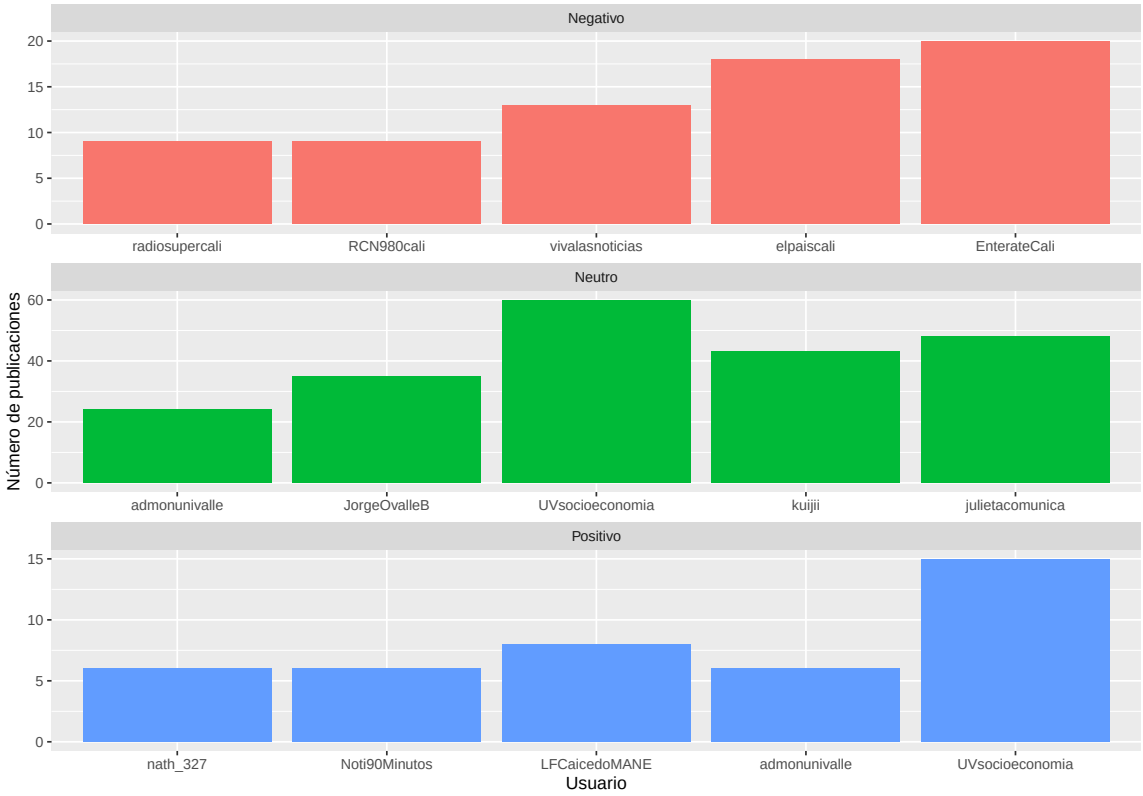


Figura 4.4: Número de publicaciones de los usuarios por sentimiento.

Capítulo 5

Conclusiones y trabajos futuros

5.1. Conclusiones

- Al evaluar los resultados que se obtuvieron en esta investigación, se puede concluir que el paro estudiantil que ocurrió en el periodo comprendido entre octubre y diciembre de 2018, no tuvo un impacto que influyera mayoritariamente de manera positiva o negativa en las opiniones de los usuarios, por el contrario la tendencia fue a tomar una posición neutral frente a la popularidad de la Universidad del Valle. De igual forma, el análisis exploratorio de los datos clasificados permitió obtener características que describen tendencias en cada sentimiento; en un caso se obtuvo que una de las palabras mas usadas en el sentimiento negativo es "alumbrado" haciendo referencia a los perjuicios que algunos estudiantes hicieron al alumbrado publico, en otro caso se obtuvo que dos de los usuarios con mayor número de publicaciones clasificadas como neutras, son "UVsocioeconomica" y "julietacomunica" los cuales sus publicaciones son mayoritariamente informativas.

- Se realizó una investigación de los métodos existentes para la obtención de las publicaciones realizadas por los usuarios de la red social Twitter en el periodo de tiempo y con los parámetros de búsqueda establecidos; el uso de web scraping facilitó esta recolección de datos, que fueron necesarios para el desarrollo de la investigación. Al mismo tiempo, el uso de diferentes técnicas de preprocesamiento de datos permitió la disminución del ruido o posibles situaciones que se puedan presentar en los datos que puedan disminuir el rendimiento de clasificación de los modelos, además de ayudar a la representación numérica de las publicaciones para el entrenamiento y evaluación de dichos modelos.

- Se analizaron cuatro de los diferentes algoritmos de aprendizaje automático que se utilizan en aplicaciones de minería de datos, los cuales fueron entrenados y evaluados con el fin de compararlos entre ellos para determinar el que obtenga los mejores resultados. Por otro lado, debido al desbalance presentado

en los datos, se aplicaron técnicas como downsampling, upsampling y smote que afectaron de manera negativa a la exactitud de los modelos, pero también de manera positiva en cuanto a la clasificaciones individuales de cada clase, medido mediante el f1 score; como resultado, se determinó que el modelo SVM aplicando la técnica de balanceo upsampling era el modelo con mejores resultados, con una exactitud de casi 60 % y con una puntuación en f1 score de más 55 % en cada clase.

- Haciendo uso del modelo con el que obtuvo los mejores resultados, se clasificaron las publicaciones obtenidas, dejando como resultado que poco más del 50 % de los casos se clasificaron con un sentimiento neutral, con lo que se concluye que, en mayor medida, las publicaciones de los usuarios de la red social Twitter tienen un sentimiento neutro hacia la Universidad del Valle.

- En el desarrollo de esta investigación, se profundizaron los conocimientos en las áreas de minería de datos y el análisis de los sentimiento, comenzando por el proceso de obtención y manipulación de los datos, pasando por el preprocesamiento de los mismos y el entrenamiento de modelos de aprendizaje automático, y la evaluación y utilización de estos modelos para el descubrimiento de conocimiento presente en los datos, lo que se convierte en un referente para adelantar futuros proyectos relacionados con el estudio de la imagen corporativa de la Universidad a través del análisis de sentimientos y otras técnicas.

5.2. Trabajos futuros

Los resultados obtenidos en esta investigación abren las puertas a varios trabajos; a continuación, se presentan algunos posibles trabajos futuros que se pueden desarrollar como resultado de esta investigación.

- Comparar los resultados obtenidos en este trabajo, con otros paros estudiantiles que tuvieran diferentes propósitos, con el fin de observar si se tiene el mismo comportamiento o, por el contrario, el propósito de los paros estudiantiles afecta en las opiniones de los usuarios.

- Implementar un corrector ortográfico con el propósito de corregir publicaciones que tengan errores ortográficos e intentar aumentar la tasa de aciertos de los modelos de clasificación.

- Ampliar el análisis de este proyecto hacia otras instituciones académicas u organizaciones, o hacia otras redes sociales con el fin de obtener mas diversidad en opiniones.

Bibliografía

[1] Hamid S., Michel L. y Marie O. Richard, "Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitte", *International Journal of Information Management*, 2017.

[2] Estévez F., García A. y Glösekötter P., "An application of people's sentiment from social media to smart cities", 2016.

[3] Gómez E. , Jaimes R., Hidalgo Orlando y Luján S., "Influencia de redes sociales en el análisis de sentimiento aplicado a la situación política en Ecuador", *Enfoque UTE*, 2016.

[4] Roales González N., "Detección de tendencias en Twitter utilizando minería de datos adaptativa" [proyecto de fin de grado], Madrid: Universidad autónoma de Madrid, Ingeniería informática, 2014.

[5] Olarte1 E., Panizzi M. y Bertone R., "Segmentación de Mercado Usando Técnicas de Minería de Datos en Redes Sociales", *XXIV Congreso Argentino de Ciencias de la Computación*, 2018.

[6] Meneses M. , Martín del Campo A. y Rueda H., "#TrumpenMéxico. Acción conectiva transnacional en Twitter y la disputa por el muro fronterizo", 2017.

[7] Bembibre C., "Definición de Paro", [Articulo de internet], https://www.definicionabc.com/social/paro.php, [Consulta: 9 febrero de 2020].

[8] Bannister K., "Entendiendo el análisis de sentimiento: qué es y para qué se usa", [Publicación periódica en linea] 2015, [citada 9 febrero de 2020], Disponible en https://www.brandwatch.com/es/blog/analisis-de-sentimiento/

[9] Riquelme J. C., Ruiz R., y Gilbert K., "Minería de Datos: Conceptos y Tendencias ", *Revista Iberoamericana de Inteligencia Artificial*, 2006, vol. 10, núm. 29, pp. 11-18.

[10] Villena Román J., "CRISP-DM: La metodología para poner orden en los proyectos", *s/ngular*, 2016.

[11] Alan R. Peslak, "Sentiment Analysis and Opinion Mining: Current State of the Art and Review of Google and Yahoo Search Engines Privacy Policies", *Penn State University*, 2017.

[12] IBM Knowledge Center, "Conceptos básicos de ayuda de CRISP-DM", 2021, Disponible en https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

[13] Halling Hilborg P. y Blædel Nygaard E., "Viability of sentiment analysis in business. Evaluating accuracy and the supporting NLP technologies", 2017.

[14] Redacción APD, "La influencia de las redes sociales en la sociedad", 2015.

[15] Balagueró T., "¿Qué son los datasets y los dataframes en el Big Data?", *Deustro formation*, 2015.

[16] He W., Wu H., Yang G., Akula V., y Shen J. "A novel social media competitive analytics framework with sentiment benchmarks", 2015, 52(7), 801-802.

[17] Jefferson-Henrique, "Get Old Tweets Programatically", A project written in Python to get old tweets, it bypass some limitations of Twitter Official API. 2018.

[18] Figuerola Carlos G., Zazo Ángel F., Rodriguez Emilio, Alonso Berrocal, y José Luis, "La Recuperación de Información en español y la normalización de términos", *Revista Iberoamericana de Inteligencia Artificial*, 2004, vol. 8, núm. 22, pp. 135-145.

[19] Microsoft., "Definición de un modelo de aprendizaje automático", 2019, Disponible en https://docs.microsoft.com/es-es/windows/ai/windows-ml/what-is-a-machine-learning-model.

[20] Packt, "SVM for churn prediction", 2021, Disponible en https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789345070/3/ch03lvl1sec30/svm-for-churn-prediction.

[21] Numerentur, ,"Árboles de Decicion - DT", Disponible en http://numerentur.org/arboles-de-decision/.

[22] Numerentur, “El algoritmo K-NN y su importancia en el modelado de datos”, Blog Mercle, 2020.

[23] Merkle, “KNN Classification using Scikit-learn”, Disponible en https://ai.plainenglish.io/knn-classification-using-scikit-learn-efb34151a8b9.

[24] IBM Knowledge Center, “El modelo de redes neuronales”, 2021.

[25] IArtificial.net, “Precision, Recall, F1, Accuracy en clasificación”, 2020.

[26] Prototipo del aplicativo web desarrollada para la investigación. https://victorruizo.shinyapps.io/app_tg/

ANEXO 1

Código fuente

El código fuente utilizado para el desarrollo de esta investigación, escrito en el lenguaje de programación R, y a través de su plataforma de desarrollo R studio, se encuentra en un repositorio en Github, disponible en el siguiente enlace:

- https://github.com/VictorRuizO/TG

En la siguiente figura se describe la composición de los elementos del repositorio, además de una breve descripción de cada uno.

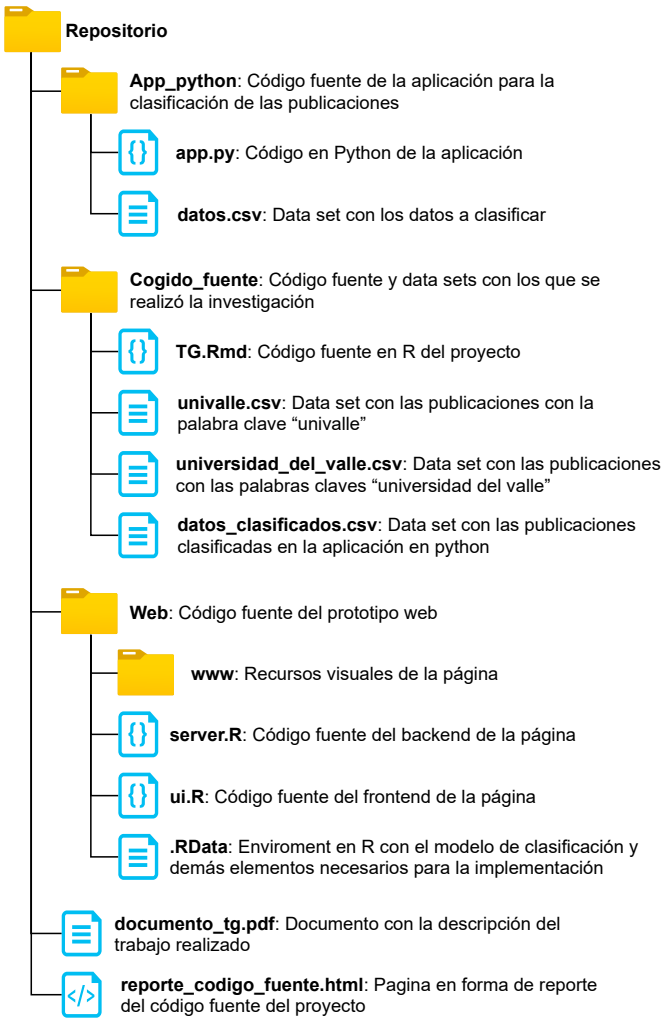


Diagrama con los elementos del repositorio.